2016 IEEE 55th Conference on Decision and Control (CDC)
ARIA Resort & Casino
December 12-14, 2016, Las Vegas, USA

Information Flow for Security in Control Systems

Sean Weerakkody Bruno Sinopoli Soummya Kar Anupam Datta

Abstract— **This paper considers the development of information flow analyses to support resilient design and active detection of adversaries in cyber physical systems (CPS). CPS security, though well studied, suffers from fragmentation. In this paper, we consider control systems as an abstraction of CPS. Here, we use information flow analysis, a well established set of methods developed in software security, to obtain a unified framework that captures and extends results in control system security. Specifically, we propose the Kullback Liebler (KL) divergence as a causal measure of information flow, which quantifies the effect of adversarial inputs on sensor outputs. We show that the proposed measure characterizes the resilience of control systems to specific attack strategies by relating the KL divergence to optimal detection. We then relate information flows to stealthy attack scenarios where an adversary can bypass detection. Finally, this article examines active detection mechanisms where a defender intelligently manipulates control inputs or the system itself to elicit information flows from an attacker's malicious behavior. In all previous cases, we demonstrate an ability to investigate and extend existing results through the proposed information flow analyses.**

I. Introduction

The security of cyber physical systems (CPS), which integrate sensing, communication, and control in physical spaces, has become a significant challenge in society [1]. Because CPS pervade our critical infrastructures including transportation, manufacturing, health care, and energy, and are often implemented using off the shelf components, they offer both motivation and opportunity for potential attackers. There exist precedence for attacks on CPS including Stuxnet [2] and the Maroochy Shire incident [3] .

The ability to detect and characterize attacks is paramount to the well being of CPS. In particular, to deliver countermeasures for attacks on physical systems, the operator must passively detect attacks in a timely manner. Moreover, the defender must understand the set of stealthy attacks to motivate resilient design and active detection. Here, passive detection refers to the defender's use of information to ascertain if the system is operating normally or under attack. Passive detection techniques against attacks in CPS have been well studied. For instance, traditional methods of fault detection [4], [5] have been considered. However, such schemes are usually designed to deal with benign failures. Consequently, recent work considered the detection of stealthy malicious adversaries who perform integrity attacks on sensor measurements and control inputs [6], [7], [8].

Despite this previous work, the detection of arbitrary attacks on CPS by adversaries with diverse information and capabilities is not well categorized. In this article, we propose using information flows as a means to quantify the detectability of generic adversarial attack models. Information flow analysis is an establised set of tools in software security [9], which determine if the processes of one agent alter the processes of another agent. We intend to use information flow to develop a unified treatment of security in CPS, specifically focusing on dynamical control aspects in this paper and leaving general cyber-physical treatments to future work.

In this article, we propose the KL divergence as a quantitative measure for information flow to determine the extent to which an attacker's inputs affect control system outputs. To complement this measure, we introduce notions of conditional ϵ-weak information flows and conditional ϵ-strong information flows. Here, weak flows characterize stealthy attack strategies conditioned on the system model and the defender's control policy. Moreover, strong flows define active defense strategies which enable attack detection, conditioned on the adversary's policy. The resulting framework allows us to recover, in a unified manner, a collection of prior results in CPS security, obtained using different techniques, in a number of papers. Moreover, in certain cases our framework allows us to present refinements on existing results to reveal additional insights. We summarize these instances below.

First, in section V, we leverage the results of [10] to show that the KL divergence characterizes optimal passive detectability by relating this measure to the optimal decay rate of the probability of false alarm. Moreover, we show through residue analysis that the KL divergence can, in many instances, be efficiently evaluated.

Next, in section VI, we consider the study of conditional weak information flows where we assume a defender chooses an arbitrary control policy. We show there exist attacks which generate 0 information flow if and only if the attacker's subsystem is not left invertible. This allows us to recover results applied by [7] to analyze undetectable attack scenarios. In addition, we show, under certain constraints on adversarial

S. Weerakkody, B. Sinopoli, S. Kar, and A. Datta are with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA 15213. Email: sweerakk@andrew.cmu.edu, brunos@ece.cmu.edu, soummyak@andrew.cmu.edu, danupam@cmu.edu.

This material is based upon work supported by the Department of Energy under Award Number DE-OE0000779. S. Weerakkody is supported in part by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program. The work of S. Weerakkody, B. Sinopoli, S. Kar, and A. Datta is supported in part by NSF grant number 1646526. The work of A. Datta is supported in part by NSF TRUST Center CCF-0424422. The work of S. Kar is supported in part by DARPA under agreement number DARPA FA8750-12-2-0291. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NSF, DARPA, DOE or the U.S. Government.

978-1-5090-1837-6/16/$31.00 ©2016 IEEE

policy, that the information flow is a quadratic function of the bias injected on measurement residues. This allow us to recover results in [11] and [12] on false data injections which used the residue bias as a constraint when studying impacts of stealthy adversaries. We are able to refine these results by presenting optimal detection guarantees for adversaries that satisfy these constraints.

Finally, information flow analysis allows us to consider results in active detection where the defender changes system parameters [13], [14], [15] or the control policy itself [16], [17], [18], [19], [20] to detect an attack. We specifically consider replay attacks. Here, we recover results which show that certain systems and control policies are vulnerable to replay attacks [16]. However, unlike [16] which uses specific continuity arguments, we use our framework to demonstrate that replay attacks generate a conditional weak information flow. We then recover results which state that introducing physical watermarking to the defender's policy [17] enables detection of replay adversaries. We do this by directly proving such a policy yields a conditional strong information flow. We are able to extend previous results [17] by using the calculated information flow to evaluate the detectability of a replay attack in a system with physical watermarking.

To close, we note that [10] also leverages results relating the KL divergence to optimal passive detectability in order to define the notion of an ϵ-stealthy attack. This is subsequently used to analyze maximum estimation degradation by a stealthy adversary in a scalar system. Our paper proposes using the KL divergence not only as a tool to analyze specific attacks, but as a unifying measure to characterize attacks and defenses in control system security. We also argue that our proposed framework is more general. Specifically, the notion of conditional information flow allows us to both characterize how an adversarial policy can be tuned to avoid detection by specific defenders and consider how the defender can adjust the system or his control policy to actively detect an attacker. We will revisit [10] in a more technical context later.

The rest of the paper is summarized as follows. In section II, we describe the system model. In section III, we introduce a general model of an adversary in a CPS. Next, in section IV we define an information flow in a CPS through the KL divergence and relate it to existing notions in software security. After, in section V, we motivate information flow as a computable measure of optimal passive detectability. In section VI, we discuss stealthy attack scenarios. Then, in section VII, we consider information flow in the context of active detection. We conclude the paper in section VIII.

II. System Model

We consider a control system with discrete linear time invariant model given below.

$$x_{k+1} = Ax_k + Bu_k + w_k, \quad y_k = Cx_k + v_k. \tag{1}$$

Here $x_k \in \mathbb{R}^n$ is the state, $u_k \in \mathbb{R}^p$ is the set of control inputs and $y_k \in \mathbb{R}^m$ is the set of sensor outputs. We let x_0 be the initial state. Furthermore, $w_k \sim \mathcal{N}(0, Q)$ and $v_k \sim \mathcal{N}(0, R)$ are independent and identically distributed (IID) process and measurement noise respectively. We consider a finite horizon up to time T. The previous linear model of a system is leveraged to derive the ensuing results related to control system security. However, we stress that the paradigm of information flows, to be introduced, can consider general nonlinear and time varying dynamical systems.

We let $\mathcal{I}_k$ be the information available to the defender at time k after making a measurement. From the defender's perspective, the initial state is unknown. However, the defender knows that $f(x_0|\mathcal{I}_{-1}) = \mathcal{N}(\hat{x}_{0|-1}, P_{0|-1})$. The defender at time -1 is aware of the system model $\mathcal{M} = \{A, B, C, Q, R, \hat{x}_{0|-1}, P_{0|-1}\}$. In total the defender's information at time k is given by

$$\mathcal{I}_k = \{y_{0:k}, u_{0:k-1}, \mathcal{M}\}. \tag{2}$$

$y_{0:k}$ refers to the finite sequence $\{y_0, \cdots, y_k\}$. Thus, the defender is a central entity having knowledge of the dynamics of the system and the history of outputs and inputs. We now define an admissible defender control strategy.

Definition 1: An admissible defender control strategy is a sequence of deterministic measureable functions $\{\mathcal{U}_0, \mathcal{U}_1, \cdots, \mathcal{U}_{T-1}\}$ where $\mathcal{U}_k : \mathcal{I}_k \rightarrow \mathbb{R}^p$ for all $k \in \{0, 1, \cdots, T-1\}$ and $u_k = \mathcal{U}_k(\mathcal{I}_k)$.

As a result, the defender computes a deterministic function of the current information to generate an input. Finally, we assume that the defender implements some passive bad data detector to determine whether the system is operating normally, denoted by a null hypothesis $\mathcal{H}_0$, or if there exist an abnormality (or possible attack), denoted by a state of $\mathcal{H}_1$. We define an admissible detector as follows.

Definition 2: An admissible defender detector strategy is a sequence of deterministic measureable functions $\{\Psi_0, \Psi_1, \cdots, \Psi_{T-1}\}$ where $\Psi_k : \mathcal{I}_k \rightarrow \{\mathcal{H}_0, \mathcal{H}_1\}$ for all $k \in \{0, 1, \cdots, T\}$.

Thus, at each time k, the defender intelligently constructs a function Ψ_k which maps the defender's available information to a decision about the state of the system.

III. Attack Model

We now introduce an adversarial environment where an attacker, depending on his capabilities and knowledge of the system, can manipulate control inputs or sensor measurements to degrade control and estimation performance. Here, we formulate an adversary's effect on a system by including additive attacker inputs u_k^a and d_k^a as follows.

$$x_{k+1} = Ax_k + Bu_k + B^a u_k^a + w_k, \tag{3}$$

$$y_k = Cx_k + D^a d_k^a + v_k. \tag{4}$$

B^a characterizes the adversarial inputs, which could be a subset of actuators the attacker usurps from the defender, or his own inputs. Without loss of generality, we assume B^a is full column rank. We assume the adversary can modify m' sensors, $\mathcal{S} = \{\gamma_1, \cdots, \gamma_{m'}\} \subseteq \{1, \cdots, m\}$. Therefore, we define $D^a \in \mathbb{R}^{m \times m'}$ entrywise as $D^a_{u,v} = \mathbf{1}_{u=\gamma_j, v=j}$.

It is assumed that $u_k^a \in \mathbb{R}^{p'}$ and $d_k^a \in \mathbb{R}^{m'}$ are unknown to the defender. Thus, a defender can only measure an adversary's effect on a system through sensor readings.

We assume at a minimum that an adversary is aware of his own attack history defined by $\{u^a_{0:k-1}, d^a_{0:k-1}\}$. Additionally, the adversary may be able to read a subset of control inputs u_k or sensor outputs y_k. For instance, if the attacker can modify channels, he may also be able to intercept signals sent along these channels, thereby utilizing a man in the middle attack. The portion of inputs and outputs the attacker and defender can read are public and are denoted u^{pu}_k, y^{pu}_k. Finally, the adversary may have some imperfect prior knowledge of the plant $\hat{\mathcal{M}}$, the controller $\hat{\mathcal{C}}$, and the detector $\hat{\mathcal{D}}$. The adversary's information is

$$\mathcal{I}^a_k = \{u^a_{0:k-1}, d^a_{0:k-1}, u^{pu}_{0:k-1}, y^{pu}_{0:k}, \hat{\mathcal{M}}, \hat{\mathcal{C}}, \hat{\mathcal{D}}\}. \qquad (5)$$

An admissible attack strategy leverages the attacker's information $\mathcal{I}^a_k$ to generate attack inputs for the system.

Definition 3: An admissible attack strategy on the plant is a sequence of deterministic measureable functions $\{\mathcal{U}^a_0, \mathcal{D}^a_0, \cdots, \mathcal{U}^a_{T-1}, \mathcal{D}^a_{T-1}, \mathcal{D}^a_T\}$ where $\mathcal{U}^a_k : \mathcal{I}^a_k \times u^{pu}_k \rightarrow \mathbb{R}^{p'}$ for all $k \in \{0, 1, \cdots, T-1\}$ and $u^a_k = \mathcal{U}^a_k(\mathcal{I}^a_k, u^{pu}_k)$. Additionally, $\mathcal{D}^a_k : \mathcal{I}^a_k \rightarrow \mathbb{R}^{m'}$ for all $k \in \{0, 1, \cdots, T\}$ and $d^a_k = \mathcal{D}^a_k(\mathcal{I}^a_k)$.

We note that while current state of the art adversarial models for control systems consider attackers who do not change their attack strategy, our model considers an attacker with the freedom to leverage all his information to construct an attack input.

IV. Information Flows in Physical Systems

In software security, an information flow exists from a private input to a public output if including the private input changes the behavior of the public output. We wish to extend this notion for adversarial inputs and sensor outputs of control systems. This section proposes a measure of information flow to characterize the detectability of adversarial strategies.

We quantify the information flow through the KL divergence between the distribution of the output under attack and the distribution of the output under normal operation [21]. For definiteness, we assume that all discrete time stochastic processes of interest considered hereafter induce (joint) distributions on the path space that are absolutely continuous with respect to Lebesgue measure. Thus, they possess densities in the usual sense. The KL divergence between a distribution with probability density function $p(x)$ and a distribution with probability density function $q(x)$ over a sample space X is given by

$$D_{KL}(p(x)||q(x)) = \int_X \log\left(\frac{p(x)}{q(x)}\right) p(x)dx. \qquad (6)$$

This definition can be generalized to probability measures [22]. The KL divergence has the following properties [21].

1) $D_{KL}(p(x)||q(x)) \geq 0$.
2) $D_{KL}(p(x)||q(x)) = 0$ if and only if $p(x) = q(x)$ almost everywhere.
3) $D_{KL}(p(x)||q(x)) \neq D_{KL}(q(x)||p(x))$.

We now use the KL divergence to define information flows in a physical system. To begin, denote the conditional distribution of the output based on apriori information as

$$\mathbb{D}^{\mathcal{M},\mathcal{U}_{0:k-1},\mathcal{U}^a_{0:k-1},\mathcal{D}^a_{0:k}}_{y_{0:k}} = f(y_{0:k}|\mathcal{I}_{-1}, \mathcal{U}_{0:k-1}, \mathcal{U}^a_{0:k-1}, \mathcal{D}^a_{0:k}).$$

Definition 4: The information flow from the attacker's inputs $(\mathcal{U}^a_{0:T-1}, \mathcal{D}^a_{0:T})$ to the defender's outputs $y_{0:T}$ is

$$IF_T = \frac{1}{T+1} D_{KL}(\mathbb{D}^{\mathcal{M},\mathcal{U}_{0:T-1},\mathcal{U}^a_{0:T-1},\mathcal{D}^a_{0:T}}_{y_{0:T}} || \mathbb{D}^{\mathcal{M},\mathcal{U}_{0:T-1},0,0}_{y_{0:T}}).$$

The proposed definition of information flows has many desirable properties, which make it compatible with existing measures of information flow in cyber security. First, the KL divergence allows us to recover the property of noninterference [23] in deterministic systems and probabilistic noninterference [24] in stochastic systems. There exists interference from a high level user to a low level user if changing high level inputs changes low level outputs.

In our model, the low level inputs are the defender's actions, the high level inputs are the attacker's actions, and the low level outputs are the defender's outputs $y_{0:k}$. In a deterministic system, if an adversary's actions change the output $y_{0:k}$, the KL divergence is infinite, reflecting the fact that there is interference. However, if the output $y_{0:k}$ is the same when the system is operating normally and under attack, indicating noninterference, the KL divergence is 0. There exists probabilistic interference from a high level user to a low level user if changing high level inputs measurably alters the distribution of low level outputs. $IF_T = 0$ if and only if there exists probabilistic noninterference.

Finally, we would like to be able to measure information flow when there exists probabilistic interference. In software security, this is done through research in quantitative information flow. A majority of previous work in software security [25] has proposed associative measures of information flow such as mutual information. Associative measures of information flow, which quantify correlation, evaluate how much information is leaked by an input to the output and thus provide utility in privacy applications.

The KL divergence however is a causal measure which directly determines how varying an attacker's inputs changes the distribution of public outputs. The extent to which an attacker's input changes the system output will mark the defender's ability to distinguish outputs under attack from outputs under normal operation and thus detect the presence of an adversary. While the software security community has begun to investigate causal measures of information flow for violation detection [26], to our knowledge, the ensuing results will be the first work applied to physical systems.

To close the section we attempt to categorize adversarial policies which generate information flows bounded above by ϵ when the defender implements a specific set of control policies or has a specific model.

Definition 5: Let $\mathbb{U}$ denote denote some fixed set of ordered pairs $(\mathcal{M}^*, \mathcal{U}^*_{0:T-1})$ consisting of models and defender control strategies. A permissible attack $(\mathcal{U}^a_{0:T-1}, \mathcal{D}^a_{0:T})$ generates a $\mathbb{U}$ conditional ϵ- weak information flow if for all $(\mathcal{U}_{0:T-1}, \mathcal{M}) \in \mathbb{U}$, $IF_T \leq \epsilon$.

Several special cases which satisfy this definition have arisen in the literature. For instance, a replay attack, generates an information flow bounded above by ϵ only for certain classes of models and strategies. Another special case is below.

Definition 6: An adversary generates a $\mathcal{M}$ conditional ϵ-weak information flow if for a specific model $\mathcal{M}$, $IF_T \leq \epsilon$, regardless of the defender's policy $\mathcal{U}_{0:T-1}$.
This special case, where we remove any constraints on the defender's policy, is equivalent to ϵ-stealthiness in [10] and contains false data injections and zero dynamic attacks which we consider in section VI. We now consider defender policies and system design which elicit information flows.

Definition 7: A change in the system $\mathcal{M}$ or a permissible control policy $\mathcal{U}_{0:T-1}$ generates a $\mathbb{U}^a$ conditional ϵ- strong information flow if for $(\mathcal{U}^a_{0:T-1}, \mathcal{D}^a_{0:T}) \in \mathbb{U}^a$, $IF_T \geq \epsilon$.
The preceding definition characterizes active detection where an adversary changes system parameters or his control policy to create an information flow. We will examine this topic further in section VII.

V. Passive Detection

In this section we motivate the KL divergence as a tool to quantify the passive detectability of an adversary and evaluate the special case of $\mathcal{M}$ conditional ϵ- weak information flows. Specifically, we show that this measure is directly related to the optimal decay rate for the probability of false alarm. We now have the following result from [10].

Theorem 8: Let $0 < \delta < 1$. Define α_k the probability of false alarm and β_k the probability of detection as follows

$$\alpha_k \triangleq \Pr\left(\Psi_k(\mathcal{I}_k) = \mathcal{H}_1 | \mathcal{H}_0\right), \ \beta_k \triangleq \Pr\left(\Psi_k(\mathcal{I}_k) = \mathcal{H}_1 | \mathcal{H}_1\right).$$

Suppose $\limsup_{k \to \infty} IF_k \geq \epsilon$. Then there exists a detector Ψ_k such that
$\beta_k \geq 1 - \delta, \ \forall k, \quad \limsup_{k \to \infty} -\frac{1}{k+1}\log(\alpha_k) \geq \epsilon.$
Alternatively, suppose additionally that the sequences generated by $y_{0:k}$ operating normally and under attack are ergodic. Suppose $\lim_{k \to \infty} IF_k \leq \epsilon$. Then for all detectors Ψ_k

$$\beta_k \geq 1 - \delta, \ \forall k \implies \limsup_{k \to \infty} -\frac{1}{k+1}\log(\alpha_k) \leq \epsilon.$$

From Theorem 8, the information flow is essentially equivalent to the optimal decay rate in the probability of false alarm and an adversary who generates an $\mathcal{M}$ conditional ϵ-weak information flow will have false alarm rate bounded above by ϵ. As a result, information flow allows us to generically evaluate and compare the detectability of different attack policies. However unlike other potential measures such as β_k, the KL divergence can be efficiently characterized.

We note that it may be difficult to compute the KL divergence of the outputs $y_{0:T-1}$ directly. For instance, if a control policy includes nonlinear feedback, the Gaussian property of the output is destroyed, likely removing the ability to obtain closed form distributions of the output. We can instead consider the normalized residue z_k, obtained from a Kalman filter [27].

$$\hat{x}_{k+1|k} = A\hat{x}_{k|k} + Bu_k, \ \hat{x}_{k|k} = (I - K_kC)\hat{x}_{k|k-1} + K_k y_k,$$

$$P_{k+1|k} = AP_{k|k-1}A^T + Q - AK_kCP_{k|k-1}A^T,$$
$$K_k = P_{k|k-1}C^T(CP_{k|k-1}C^T + R)^{-1}, \tag{7}$$
$$z_k = (CP_{k|k-1}C^T + R)^{-\frac{1}{2}}(y_k - C\hat{x}_{k|k-1}). \tag{8}$$

The Kalman filter computes optimal state estimates $\hat{x}_{k|k-1}$ and $\hat{x}_k$ of x_k. The normalized residue z_k is a normalized measure of the difference between the defender's outputs and the expected outputs derived from the state estimate.

Lemma 9: [28] $f(z_{0:k}|\mathcal{I}_{-1}) = \mathcal{N}(0, I)$ when the system is operating normally. Given strategy $\mathcal{U}_{0:k-1}$ and $\hat{x}_{0|-1}$, $z_{0:k}$ is an invertible function of $y_{0:k}$.
Because the residues and outputs are related by an invertible mapping, we can show their KL divergences are equal [22].

Theorem 10: The KL divergence between sensor outputs and between residues are equivalent.

$$D_{KL}(\mathbb{D}_{y_{0:T}}^{\mathcal{M},\mathcal{U}_{0:T-1},\mathcal{U}^a_{0:T-1},\mathcal{D}^a_{0:T}} || \mathbb{D}_{y_{0:T}}^{\mathcal{M},\mathcal{U}_{0:T-1},0,0})$$
$$= D_{KL}(\mathbb{D}_{z_{0:T}}^{\mathcal{M},\mathcal{U}_{0:T-1},\mathcal{U}^a_{0:T-1},\mathcal{D}^a_{0:T}} || \mathbb{D}_{z_{0:T}}^{\mathcal{M},\mathcal{U}_{0:T-1},0,0}).$$

Due to theorem 10, we can analyze the residues operating normally and under attack instead of the system output when computing the information flow. Residues under normal operation have a known zero-mean Gaussian distribution. If the distribution of the residue under attack remains Gaussian, a closed form solution exists for the KL divergence. The KL divergence between two Gaussian distributions $\mathcal{N}_1 = \mathcal{N}_1(\mu_1, \Sigma_1)$ and $\mathcal{N}_0 = \mathcal{N}_0(\mu_0, \Sigma_0)$ with $\mu_1 \in \mathbb{R}^l$ is [21]

$$D_{KL}(\mathcal{N}_1 || \mathcal{N}_0) = -\frac{l}{2} + \frac{1}{2}\text{tr}(\Sigma_0^{-1}\Sigma_1) + \frac{1}{2}\log\det\left(\Sigma_0\Sigma_1^{-1}\right)$$
$$+ \frac{1}{2}(\mu_1 - \mu_0)^T\Sigma_0^{-1}(\mu_1 - \mu_0). \tag{9}$$

If the attacker's policy is independent of the defender's outputs, it is known that the distribution of residues under attack remain Gaussian. In general however, it may still be difficult to compute the KL divergence of $z_{0:k}$ since it is a growing sequence. Fortunately, we can leverage the independence of the residues to obtain the following bound.

Theorem 11: The information flow generated by an adversary can be lower bounded by the sum of the residue-based KL divergences generated at each time step.

$$IF_T \geq \sum_{k=0}^{T} \frac{D_{KL}(\mathbb{D}_{z_k}^{\mathcal{M},\mathcal{U}_{0:k-1},\mathcal{U}^a_{0:k-1},\mathcal{D}^a_{0:k}} || \mathbb{D}_{z_k}^{\mathcal{M},\mathcal{U}_{0:k-1},0,0})}{T+1}.$$

Proof:
By Theorem 10 and Bayes rule we know

$$IF_T = \sum_{k=0}^{T} \frac{D_{KL}(\mathbb{D}_{z_k|z_{0:k-1}}^{\mathcal{M},\mathcal{U}_{0:k-1},\mathcal{U}^a_{0:k-1},\mathcal{D}^a_{0:k}} || \mathbb{D}_{z_k}^{\mathcal{M},\mathcal{U}_{0:k-1},0,0})}{T+1}.$$

Thus, we observe

$$IF_T - IF_T^{LB} = \sum_{k=0}^{T} \frac{I_{z_k, z_{0:k-1}}^{\mathcal{M},\mathcal{U}_{0:k-1},\mathcal{U}^a_{0:k-1},\mathcal{D}^a_{0:k}}}{k+1}.$$

where IF_T^{LB} is the obtained lower bound and $I_{z_k, z_{0:k-1}}$ is the mutual information [21] which is nonnegative. ■

Instead of computing the KL divergence of vectors $z_{0:k} \in R^{mk}$, we can instead obtain a recursive lower bound by computing the sum of T divergences for vectors $z_k \in \mathbb{R}^m$. Also, the gap between the lower bound and IF_T is the scaled sum of mutual informations between z_k and $z_{0:k-1}$ so that if attack residues are independent, the gap is 0.

VI. Stealthy Adversarial Behavior

We next describe attacks which generate $\mathcal{M}$ conditional ϵ-weak information flows, where regardless of the defender's policy the attacker remains stealthy. Understanding these scenarios motivate resilient design of $\mathcal{M}$ and also allow us to capture and extend research on left invertibility and false data injection attacks. The first scenario we consider is when $\epsilon = 0$ where there exists probabilistic noninterference.

Let $y^a_{0:T}$ denote outputs realized from the distribution under attack $\mathbb{D}^{\mathcal{M},\mathcal{U}_{0:T-1},\mathcal{U}^a_{0:T-1},\mathcal{D}^a_{0:T}}_{y_{0:T}}$ and $y_{0:T}$ denote outputs realized from the normal system $\mathbb{D}^{\mathcal{M},\mathcal{U}_{0:T-1},0,0}_{y_{0:T}}$. If $\mathcal{U}_{0:T-1} = 0$, then, due to the linearity of our model $\mathcal{M}$,

$$y^a_{0:T} = y_{0:T} + \Delta y_{0:T}(d^a_{0:T}, u^a_{0:T-1}), \tag{10}$$

$$\Delta x_{k+1} = A\Delta x_k + B^a u^a_k,\ \Delta x_0 = 0, \tag{11}$$

$$\Delta y_k = C\Delta x_k + D^a d^a_k. \tag{12}$$

We now obtain the following result.

Theorem 12: A nonzero attack strategy $(\mathcal{U}^a_{0:T-1}, \mathcal{D}^a_{0:T})$ generates a $\mathcal{M}$ conditional 0-weak information flow if and only if $\Delta y_{0:T}(d^a_{0:T}, u^a_{0:T-1}) = 0$ with probability 1.

Proof: Suppose $\Delta y_{0:T}(d^a_{0:T}, u^a_{0:T-1}) = 0$ with probability $1-\epsilon$ where $\epsilon > 0$. Then for $\mathcal{U}_{0:T-1} = 0$, we have with probability $1-\epsilon$, $y^a_{0:T} \neq y_{0:T}$. Thus, the KL divergence is greater than 0. Now instead suppose $\Delta y_{0:T}(d^a_{0:T}, u^a_{0:T-1}) = 0$ with probability 1. From (3) and (4), we observe that (10) holds if $\Delta y_{0:T}(d^a_{0:T}, u^a_{0:T-1}) = 0$. This is based on the fact that the defender's control strategy will not change if the output does not change. Thus, if $\Delta y_{0:T}(d^a_{0:T}, u^a_{0:T-1}) = 0$ with probability 1, then $y^a_{0:T} = y_{0:T}$ with probability 1. Therefore, the KL divergence and information flow is 0. ■

We have shown that there exists a 0-information flow attack if and only if there exists nontrivial $(\mathcal{U}^a_{0:T-1}, \mathcal{D}^a_{0:T})$ which satisfy (11), (12) for $0 \leq k \leq T$. For long enough time horizon this is in fact equivalent to left invertibility.

Theorem 13: Let $\hat{B}^a = \begin{bmatrix} B^a & 0_{n\times m'} \end{bmatrix}$, $\hat{D}^a = \begin{bmatrix} 0_{m\times p'} & D^a \end{bmatrix}$. Suppose $T \geq n - p' + 1$. A nonzero adversarial policy $(\mathcal{U}^a_{0:T-1}, \mathcal{D}^a_{0:T})$ can generate a $\mathcal{M}$ conditional 0-weak information flow if and only if $(A, \hat{B}^a, C, \hat{D}^a)$ is not left invertible.

Proof: The result follows directly from Theorem 12 and Corollary 1 of [29]. ■

Left invertibility in control systems has been well studied in previous work in CPS security as a subset of zero dynamic attacks [7]. Our general framework of information flows is able to recover this property and consequently, we can directly apply previous results related to left invertibility in our study of 0-weak information flows. For instance, we can consider conditions on $\mathcal{M}$ which allow for the existence of 0 information flow attacks to motivate resilient design of the system (A, B, C) and channel security (B^a, D^a).

Theorem 14: [7] Let $T \geq n - p' + 1$. An attack policy can create a $\mathcal{M}$ conditional 0-weak information flow if and only if $\text{rank}\left(\bar{P}(\mathcal{M})\right) < n + p' + m', \quad \forall\, \lambda \in \mathbb{C}$

$$\text{where } \bar{P}(\mathcal{M}) = \begin{bmatrix} \lambda I - A & \hat{B}^a \\ C & \hat{D}^a \end{bmatrix}.$$

We now wish to consider the case of $\mathcal{M}$ conditional ϵ-weak information flows for $\epsilon > 0$. However, we assume that the adversary injects additive inputs which are independent of the defender's system inputs and outputs. Thus, we assume

$$\begin{aligned} u^a_k &= \mathcal{U}^a_k(u^a_{0:k-1}, d^a_{0:k}, \hat{\mathcal{M}}, \hat{\mathcal{C}}, \hat{\mathcal{D}}), \\ d^a_k &= \mathcal{D}^a_k(u^a_{0:k-1}, d^a_{0:k-1}, \hat{\mathcal{M}}, \hat{\mathcal{C}}, \hat{\mathcal{D}}). \end{aligned} \tag{13}$$

Such attacks are known as false data injection attacks.

Theorem 15: Consider an admissible adversarial policy which satisfies (13). Then,

$$IF_T = \frac{1}{2(T+1)} \Delta z^T_{0:T} \Delta z_{0:T}, \tag{14}$$

where Δz_k satisfies $\Delta e_{0|-1} = 0$ and

$$\begin{aligned} \Delta e_{k+1|k} &= (A - AK_kC)\Delta e_{k|k-1} + B^a u^a_k - AK_k D^a d^a_k, \\ \Delta z_k &= (CP_{k|k-1}C^T + R)^{-\frac{1}{2}} \left(C\Delta e_{k|k-1} + D^a d^a_k\right). \end{aligned} \tag{15}$$

Proof: See Appendix I. ■

Thus, the information flow is proportional to the norm of Δz_k squared where Δz_k represents the bias the adversary injects on the normalized residue. The norm of the residue bias has been previously used as a measure of the stealthiness in false data injection attacks. For instance, [11] and [12], in their investigation of false data injection attacks, restrict

$$\|\Delta z_k\|^2 \leq M \quad \forall k. \tag{16}$$

with the motivation that the increase in β_k will be bounded by some M' in this scenario. For $M \leq 2\epsilon$, such an attacker generates a $\mathcal{M}$ conditional ϵ-weak information flow. Consequently we have the following result.

Theorem 16: Suppose a false data injection attack satisfies $\|\Delta z_k\|^2 \leq 2\epsilon \quad \forall k$. Then, for $\delta > 0$ there exists a detector such that $\beta_k \geq 1 - \delta$ and $\limsup\limits_{k\to\infty} -\frac{\log(\alpha_k)}{k+1} = \epsilon$.

Again, the results obtained in [11], evaluating models $\mathcal{M}$ and attacks $\mathcal{D}^a_{0:k}$ which stealthily destabilize a system, and [12], estimating the bias an adversary can stealthily inject on the system state in $\mathcal{M}$, can all be reframed as attacks which generate $\mathcal{M}$ conditional ϵ-weak information flow. This refinement of existing results allows us to now quantify detectability in addition to system impact.

VII. Active Detection of Adversarial Behavior

In this section, we will revisit and extend results related to the active detection of replay attacks using the proposed measure of information flow. Recall that in active detection, the defender changes the system or his policy to elicit an information flow. Specifically, we will use information flows to determine when replay attacks are stealthy. We will then extend previous work by using information flows to characterize optimal detection with watermarking.

In a replay attack, the adversary observes a sequence of measurements from y_{-N} to y_{-N+T-1}. Then, without loss of generality, at time 0, the attacker replays these measurements. Here, we will assume $-N$ is large so that the adversary has an adequate buffer and that the replayed outputs are independent of the current outputs. Moreover, we assume the system at time $-N$ is in steady state. We first argue that a replay attack generates a $\mathbb{U}$ conditional ϵ-weak information flow for a large class of systems and common control policies. For instance, consider a defender that uses state feedback with gain L so $\mathcal{U}_k(\mathcal{I}_k) = L\hat{x}_{k|k}$.

Let $\mathcal{A} = (A + BL)(I - KC)$ and $\mathcal{P} = CPC^T + R$. It has been shown that [17]

$$z_k = z_{k-N} - \mathcal{P}^{-\frac{1}{2}} C\mathcal{A}^k(\hat{x}_{0|-1} - \hat{x}_{-N|-N-1}). \qquad (17)$$

If $\mathcal{M}$ and $\mathcal{U}_{0:k-1}$ generate stable $\mathcal{A}$ the second term converges to 0. Therefore, we have the following result regarding the information flow with proof in appendix II.

Theorem 17: Suppose that our control system (1) with state feedback control is under replay attack, where $\rho(\mathcal{A}) < 1$. Then, $\lim\limits_{T\to\infty} IF_T = 0$.

If $\mathcal{A}(\mathcal{M}, \mathcal{U}_{0:k-1})$ is stable, the adversary's actions are asymptotically undetectable since the information flow is 0. This result was previously obtained in [16] by instead showing that continuous functions of the defender's information are indistinguishable under normal and replay scenarios. Information flows allow us to recover this result via a general CPS security framework.

In this example, the defender's control strategy $\mathcal{U}_{0:T-1}$ of state feedback, leaves the system vulnerable to a replay attack. The defender ideally should be able to perform active detection and determine a control strategy which simultaneously addresses system objectives while creating an information flow from a replay adversary.

Watermarking techniques allow the defender to increase the information flow from the attacker input to defender output and as a result create an $\mathbb{U}^a$ conditional ϵ-strong information flow, where $\mathbb{U}^a$ contains the replay attack policy. In watermarking, noisy control inputs are used with $u_k = \mathcal{U}_k(\mathcal{I}_k) = L\hat{x}_{k|k} + \Delta u_k$ where $\Delta u_k \sim \mathcal{N}(0, \mathcal{Q})$. Note that while the watermark is random, it can be predetermined offline so that $\mathcal{U}_k(\mathcal{I}_k)$ remains a deterministic function. We now show watermarking creates a strong information flow.

Theorem 18: Suppose the system (1) with state feedback control and watermarking is under replay attack, where $\rho(\mathcal{A}) < 1$. Then, almost surely $\lim\limits_{T\to\infty} IF_T \geq \epsilon$, where

$$\epsilon = \frac{\text{tr}\left(\mathcal{P}^{-1}C\Sigma C^T\right)}{2}, \quad \Sigma = \mathcal{A}\Sigma\mathcal{A}^T + B\mathcal{Q}B^T.$$

Proof: See Appendix III. ■

From the theorem above, the defender can make the information flow from an adversarial input arbitrarily large by increasing $\text{tr}\left(\mathcal{P}^{-1}C\Sigma C^T\right)$ which is a linear function of the watermark covariance $\mathcal{Q}$. In fact, previous work on watermarking [17] does aim to design watermarks by maximizing $\text{tr}\left(\mathcal{P}^{-1}C\Sigma C^T\right)$ subject to constraints on control performance in the system. Thus, our results motivate the choice of this objective function. The use of information flows also allows us to extend previous results to analyze optimal detection of replay attacks under watermarking scenarios.

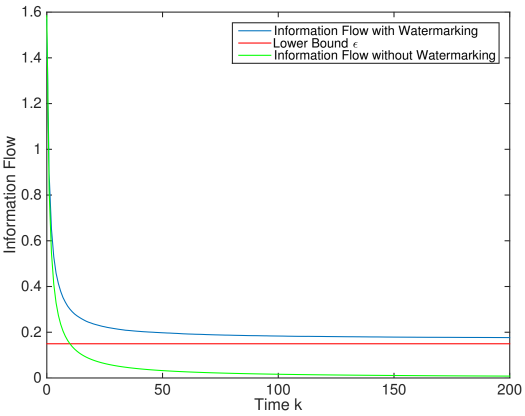

Fig. 1. Information Flow generated by a replay attack. The information flow as a function of k in the presence of watermarking is included along with its lower bound ϵ, and the information flow generated when physical watermarking is not present

Corollary 19: Assume system (1) with state feedback control and watermarking is under replay attack, where $\rho(\mathcal{A}) < 1$. Then for $\delta > 0$ there exists a detector such that $\beta_k \geq 1 - \delta$, $\forall\ k$ and

$$\limsup_{k\to\infty} -\frac{1}{k+1}\log(\alpha_k) \geq \frac{\text{tr}\left(\mathcal{P}^{-1}C\Sigma C^T\right)}{2}. \qquad (18)$$

Proof: The result follows from Theorems 18 and 8. ■

We simulate a vehicle moving along a single axis [11] under replay attack. Here, we assume that the defender obtains the gain L using a linear quadratic Gaussian (LQG) controller which attempts minimize a cost J given by

$$J = \lim_{T\to\infty} \frac{1}{T+1}\mathbb{E}\left[\sum_{k=0}^{T} x_k^T x_k + u_k^T u_k\right].$$

The LQG cost increases linearly with $\mathcal{Q}$. We select the covariance $\mathcal{Q}$ of the watermark so that ΔJ, the increased cost due to watermarking, is 40% of the optimal J. Here, we simulate the system 1000 times over a horizon of 200 steps. We plot the average information flow in Fig 1, both with watermarking and without watermarking. As expected from Theorem 17, in the absence of watermarking, the information flow generated by a replay attack converges to 0. If physical watermarking is implemented, the information flow generated by an adversary has a lower bound ϵ which grows linearly with $\mathcal{Q}$. We implement a Neyman Pearson detector [21] and plot the average probability of false alarm and detection as a function of k in Fig 2.

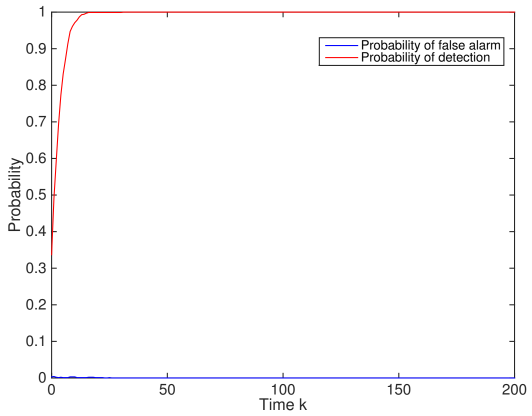

Fig. 2. Probability of detection and probability of false alarm vs time for a Neyman Pearson Detector

VIII. Conclusion

In this article, we introduced a physical measure of information flow to characterize detection in CPS and provide a unified approach to dealing with security in both the cyber and physical domains. We proposed the KL divergence as a measure of information flow. We motivate its use through results in optimal passive detection and computational ease of evaluation. We examined attacks which are stealthy for fixed models, and all input strategies, recovering results related to left invertibility and false data injection attacks. Finally, we investigated replay attacks and used information flows to quantify optimal detection performance with physical watermarking. We close by noting that information flow tools are amenable to true CPS analysis. In particular, we can consider a richer set of problems emcompassing both cyber and physical domains by leveraging the proposed results in physical security and existing parallels in cyber and software security. Approaching these general problems will mark the next stage of obtaining a unified paradigm for addressing CPS security.

Appendix I
Proof of Theorem 15

Proof: Let $e_{k|k-1} = x_k - \hat{x}_{k|k-1}$. From (3),(4), and (7), we obtain

$$e_{k+1|k} = (A - AK_kC)e_{k|k-1} + B^a u_k^a + w_k - AK_k v_k - AK_k D^a d_k^a,$$

$$z_k = (CP_{k|k-1}C^T + R)^{-\frac{1}{2}} \left(Ce_{k|k-1} + v_k + D^a d_k^a \right).$$

Let z_k^s be the residue under normal operation. Then,

$$e_{k+1|k}^s = (A - AK_kC)e_{k|k-1}^s + w_k - AK_k v_k, \; e_{0|-1}^s = e_{0|-1}$$

$$z_k^s = (CP_{k|k-1}C^T + R)^{-\frac{1}{2}} \left(Ce_{k|k-1}^s + v_k \right).$$

It can be seen from the linearity of the system that

$$z_k = z_k^s + \Delta z_k,$$

and that (15) holds. Moreover, from an inductive argument, we see that Δz_k is a deterministic variable since $\mathcal{U}_{0:k-1}^a$ and $\mathcal{D}_{0:k}^a$ are independent of $y_{0:k}$. As a result, $\mathbb{D}_{z_{0:k}}^{\mathcal{M},\mathcal{U}_{0:k-1},\mathcal{U}_{0:k-1}^a,\mathcal{D}_{0:k}^a} = \mathcal{N}(\Delta z_{0:k}, I)$. Finally, from (9) and Theorem 15, we have

$$D_{KL}\left(\mathcal{N}(\mu_1, \Sigma_1), \mathcal{N}(\mu_2, \Sigma_1)\right) = \frac{1}{2}\|\Sigma_1^{-\frac{1}{2}}(\mu_1 - \mu_2)\|^2.$$

The result immediately follows. ■

Appendix II
Proof of Theorem 17

Proof: We observe from (17) that

$$z_{0:k} \sim \mathcal{N}(\mu_r, \Sigma_r), \tag{19}$$

$$\mu_r(jm : jm + m - 1) = \mathbb{E}[z_j] = -\mathcal{P}^{-\frac{1}{2}} C \mathcal{A}^k \hat{x}_{0|-1}, \tag{20}$$

$$\begin{aligned} \Sigma_r(jm : jm + m - 1, lm : lm + m - 1) &= \text{Cov}(z_j, z_l^T), \\ &= \mathcal{P}^{-\frac{1}{2}} C \mathcal{A}^j \mathcal{W} (\mathcal{A}^l)^T C^T \mathcal{P}^{-\frac{1}{2}} + \delta(l - m) I, \end{aligned} \tag{21}$$

where $\mathcal{W}$ is the steady state covariance of $\hat{x}_{k|k-1}$ and δ refers to the discrete delta dirac function. From (9), Theorem 10, and Sylvester's determinant theorem we have

$$D_{KL}(\mathbb{D}_{y_{0:k}}^{\mathcal{M},\mathcal{U}_{0:k-1},\mathcal{U}_{0:k-1}^a,\mathcal{D}_{0:k}^a} || \mathbb{D}_{y_{0:k}}^{\mathcal{M},\mathcal{U}_{0:k-1},0,0}) = \frac{c_1 + c_2 + c_3}{2}$$

where

$$c_1 = \text{tr}\left(\sum_{j=0}^{k} \mathcal{P}^{-\frac{1}{2}} C \mathcal{A}^j \mathcal{W} (\mathcal{A}^j)^T C^T \mathcal{P}^{-\frac{1}{2}} \right),$$

$$c_2 = \sum_{j=0}^{k} \hat{x}_{0|-1}^T (\mathcal{A}^j)^T C^T \mathcal{P}^{-1} C \mathcal{A}^j \hat{x}_{0|-1},$$

$$c_3 = -\log\det\left(I + \sum_{j=0}^{k} \mathcal{W}^{\frac{1}{2}} (\mathcal{A}^j)^T C^T \mathcal{P}^{-1} C \mathcal{A}^j \mathcal{W}^{\frac{1}{2}} \right).$$

Let X_1 and X_2 be given by

$$X_1 = \sum_{j=0}^{\infty} \mathcal{A}^j \mathcal{W} (\mathcal{A}^j)^T = \mathcal{A} X_1 \mathcal{A}^T + W,$$

$$X_2 = \sum_{j=0}^{\infty} (\mathcal{A}^j)^T C^T \mathcal{P}^{-1} C \mathcal{A}^j = \mathcal{A}^T X_2 \mathcal{A} + C^T \mathcal{P}^{-1} C.$$

From Lyapunov's equation and since $\mathcal{A}$ is stable, the matrices X_1 and X_2 exist and are bounded. Since c_1, c_2, and $|c_3|$ are monotonic in k, we have for all k

$$c_1 \leq \text{tr}\left(\mathcal{P}^{-\frac{1}{2}} C X_1 C^T \mathcal{P}^{-\frac{1}{2}} \right), \; c_2 \leq \hat{x}_{0|-1}^T X_2 \hat{x}_{0|-1}^T,$$

$$|c_3| \leq \log\det\left(I + \mathcal{W}^{\frac{1}{2}} X_2 \mathcal{W}^{\frac{1}{2}} \right).$$

Consequently, for all k there exists M^* satisfying

$$D_{KL}(\mathbb{D}_{y_{0:k}}^{\mathcal{M},\mathcal{U}_{0:k-1},\mathcal{U}_{0:k-1}^a,\mathcal{D}_{0:k}^a} || \mathbb{D}_{y_{0:k}}^{\mathcal{M},\mathcal{U}_{0:k-1},0,0}) \leq M^*,$$

Dividing by $k + 1$, the result follows. ■

Appendix III
Proof of Theorem 24

Proof: When under a replay attack, we have [17]

$$z_k = z_{k-N} - \mathcal{P}^{-\frac{1}{2}} C \mathcal{A}^k (\hat{x}_{0|-1} - \hat{x}_{-N|-N-1}) - \mathcal{P}^{-\frac{1}{2}} C \sum_{j=0}^{k-1} \mathcal{A}^{k-1-j} B \left(\Delta u_j - \Delta u_{j-N}\right), \tag{22}$$

where N is some unknown, but large delay between the replayed sequence and the true sequence. Thus, under attack $z_k \sim \mathcal{N}(\mu_k, \Sigma_k + I)$ with

$$\mu_k = \mathcal{P}^{-\frac{1}{2}} C \mathcal{A}^k \hat{x}_{0|-1} + \mathcal{P}^{-\frac{1}{2}} C \sum_{j=0}^{k-1} \mathcal{A}^{k-1-j} B \Delta u_j,$$

$$\Sigma_k = \mathcal{P}^{-\frac{1}{2}} C [\mathcal{A}^k W \mathcal{A}^{k\ T} + \sum_{j=0}^{k-1} \mathcal{A}^j B \mathcal{Q} B^T \mathcal{A}^{j\ T}] C^T \mathcal{P}^{-\frac{1}{2}}.$$

Thus, the KL divergence between z_k under attack and under normal operation is given by

$$D_{KL}(\mathbb{D}_{z_k}^{\mathcal{M}, \mathcal{U}_{0:k-1}, \mathcal{U}^a_{0:k-1}, \mathcal{D}^a_{0:k}} || \mathbb{D}_{z_k}^{\mathcal{M}, \mathcal{U}_{0:k-1}, 0, 0}) = \frac{c_k^1 + c_k^2 + c_k^3}{2} \tag{23}$$

where

$$c_k^1 = \mu_k^T \mu_k, \quad c_k^2 = -\log\det\left(I + \Sigma_k\right), \quad c_k^3 = \mathrm{tr}(\Sigma_k).$$

From [19], it is known that

$$c_k^2 + c_k^3 \geq 0. \tag{24}$$

Furthermore, by the law of large numbers, we know

$$\lim_{T \to \infty} \frac{1}{T+1} \sum_{k=0}^{T} c_k^1 \overset{a.s.}{\to} \mathrm{tr}\left(\mathcal{P}^{-1} C \Sigma C^T\right). \tag{25}$$

Using (23), (24) and (25)

$$\lim_{T \to \infty} \sum_{k=0}^{T} \frac{D_{KL}(\mathbb{D}_{z_k}^{\mathcal{M}, \mathcal{U}_{0:k-1}, \mathcal{U}^a_{0:k-1}, \mathcal{D}^a_{0:k}} || \mathbb{D}_{z_k}^{\mathcal{M}, \mathcal{U}_{0:k-1}, 0, 0})}{T+1} \geq \epsilon. \tag{26}$$

By Theorem 11, the result immediately follows. ■

References

[1] T. Cardenas, A. A.and Roosta and S. Sastry, "Rethinking security properties, threat models, and the design space in sensor networks: A case study in scada systems," *Ad Hoc Networks*, vol. 7, no. 8, pp. 1434–1447, 2009.

[2] R. Langner, "To kill a centrifuge: A technical analysis of what stuxnet's creators tried to achieve," Langner Communications, Tech. Rep., November 2013. [Online]. Available: www.langner.com/en/wp-content/uploads/2013/11/To-kill-a-centrifuge.pdf

[3] J. Slay and M. Miller, "Lessons learned from the maroochy water breach," in *Critical Infrastructure Protection*. Springer US, 2008, pp. 73–82.

[4] H. L. Jones, "Failure detection in linear systems," Ph.D. dissertation, M.I.T., Cambridge, Massachusetts, 1973.

[5] A. S. Willsky, "A survey of design methods for failure detection in dynamic systems," *Automatica*, vol. 12, pp. 601–611, Nov 1976.

[6] Y. Mo, J. Hespanha, and B. Sinopoli, "Robust detection in the presence of integrity attacks," in *American Control Conference (ACC), 2012*, June 2012, pp. 3541–3546.

[7] F. Pasqualetti, F. Dorfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *Automatic Control, IEEE Transactions on*, vol. 58, no. 11, pp. 2715–2729, Nov 2013.

[8] S. Sundaram, M. Pajic, C. Hadjicostis, R. Mangharam, and G. J. Pappas, "The wireless control network: monitoring for malicious behavior," in *IEEE Conference on Decision and Contro*, Atlanta, GA, Dec 2010.

[9] D. E. Denning and P. J. Denning, "Certification of programs for secure information flow," *Commun. ACM*, vol. 20, no. 7, pp. 504–513, 1977. [Online]. Available: http://doi.acm.org/10.1145/359636.359712

[10] C.-Z. Bai, F. Pasqualetti, and V. Gupta, "Security in stochastic control systems: Fundamental limitations and performance bounds," in *American Control Conference (ACC), 2015*, June 2015.

[11] Y. Mo and B. Sinopoli, "False data injection attacks in control systems," in *First Workshop on Secure Control Systems*, Stockholm, Sweden, April 2010.

[12] ——, "Integrity attacks on cyber-physical systems," in *Proceedings of the 1st international conference on High Confidence Networked Systems*. ACM, 2012, pp. 47–54.

[13] A. Teixeira, I. Shames, H. Sandberg, and K. Johansson, "Revealing stealthy attacks in control systems," in *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*, Oct 2012, pp. 1806–1813.

[14] F. Miao, Q. Zhu, M. Pajic, and G. Pappas, "Coding sensor outputs for injection attacks detection," in *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*, Dec 2014, pp. 5776–5781.

[15] S. Weerakkody and S. B., "Detecting integrity attacks on control systems using a moving target approach," in *Submitted to Decision and Control (CDC), 2015 IEEE 54th Annual Conference on*, Dec 2015.

[16] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *Communication, Control, and Computing, 2009. Allerton 2009. 47th Annual Allerton Conference on*, Sept 2009, pp. 911–918.

[17] Y. Mo, R. Chabukswar, and B. Sinopoli, "Detecting integrity attacks on scada systems," *Control Systems Technology, IEEE Transactions on*, vol. 22, no. 4, pp. 1396–1407, July 2014.

[18] S. Weerakkody, Y. Mo, and B. Sinopoli, "Detecting integrity attacks on control systems using robust physical watermarking," in *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*, Dec 2014, pp. 3757–3764.

[19] Y. Mo, S. Weerakkody, and B. Sinopoli, "Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs," *Control Systems, IEEE*, vol. 35, no. 1, pp. 93–109, Feb 2015.

[20] F. Miao, M. Pajic, and G. Pappas, "Stochastic game approach for replay attack detection," in *Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on*, Dec 2013, pp. 1854–1859.

[21] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

[22] S. Kullback, *Information theory and statistics*. Courier Corporation, 1968.

[23] J. A. Goguen and J. Meseguer, "Security policies and security models," in *IEEE Symposium on Security and Privacy*, 1982, pp. 11–20.

[24] D. M. Volpano and G. Smith, "Probabilistic noninterference in a concurrent language," *Journal of Computer Security*, vol. 7, no. 1, 1999.

[25] G. Smith, "On the foundations of quantitative information flow," in *Foundations of Software Science and Computational Structures, 12th International Conference, FOSSACS 2009, Held as Part of the Joint European Conferences on Theory and Practice of Software, ETAPS 2009, York, UK, March 22-29, 2009. Proceedings*, 2009, pp. 288–302.

[26] A. Datta, S. Sen, and Y. Zick, "Algorithmic transparency via quantitative input influence," in *37th IEEE Symposium on Security and Privacy*, 2016.

[27] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Fluids Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[28] R. K. Mehra and J. Peschon, "An innovations approach to fault detection and diagnosis in dynamic systems," *Automatica*, vol. 7, no. 5, pp. 637–640, 1971.

[29] A. S. Willsky, "On the invertibility of linear systems," *IEEE Transactions on Automatic Control*, vol. 19, no. 3, pp. 272–274, 1974.